Procedia Computer Science

Volume 88, 2016, Pages 107–114

7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

A Framework based on Semantic Spaces and Glyphs for Social Sensing on Twitter

Giovanni Pilato[1] and Umberto Maniscalco[1]

ICAR-CNR, Italian National Research Council Palermo, Italy
{giovanni.pilato,umberto.maniscalco}@cnr.it

Abstract

In this paper we present a framework aimed at detecting emotions and sentiments in a Twitter stream. The approach uses the well-founded Latent Semantic Analysis technique, which can be seen as a bio-insipred cognitive architecture, to induce a semantic space where tweets are mapped and analysed by soft sensors. The measurements of the soft sensors are then used by a visualisation module which exploits glyphs to graphically present them. The result is an interactive map which makes easy the exploration of reactions and opinions in the whole globe regarding tweets retrieved from specific queries.

Keywords: Social Sensing, LSA, Soft Sensors, Emotion analysis, Sentiment Analysis

1 Introduction

Microblogging has shown a massive growth in popularity in recent years. One of the most representative examples of microblogging services is Twitter [19]. On the other hand, the interest in social sensing, named also participatory sensing [1],[6], together with sentiment analysis methods and techniques for recognition of emotions and opinions in social networks have become a field of great interest. Since sentiments and emotions characterize many aspects of human living many researchers tried to identify if a text can be considered as being subjective or objective [2] and also if an opinion expressed has a positive or negative polarity [14][13]. Moreover, understanding the emotional content expressed in social networks can help in inferring and describing the emotional status of a community, a group of people, a city, or even a country with regard to specific topics[16][9].

In this paper we illustrate a preliminary framework which tries to implement social sensing on Twitter by exploiting a semantic space that can be automatically induced from data. The semantic space is built by means of a Latent Semantic Analysis (LSA) [11] approach applied to a set of tweets downloaded from the Twitter streaming. The LSA approach has been chosen since it has been successfully used to simulate many psycholinguistic phenomena [12] and it has also a statistical foundation[15]. Once the semantic space is built, it can be exploited to map in it the tweets that are sensed from the Twitter stream according to given queries. The system

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.413

uses three soft sensors on the "perceived" tweets trying to infer both sentiments and emotions expressed in them, according to the belief that "it is possible to infer emotion properties from the emotion words"[5]. The first soft sensor is aimed at recognizing the sentiment orientation (positive, neutral or negative) in a tweet; the second one is aimed at sensing the emotions in a tweet. A third soft sensor extracts the geospatial coordinates, if they are available, of the tweets under analysis. The general emotional content regarding the topics described by the keywords is modeled and specified on Ekman's emotional scale [7], which assumes that there is a finite number of basic and discrete emotions and specifies the following six human emotions: anger, disgust, fear, happiness, sadness and surprise [3][4]. Once the sensing is completed, the information coming from the two sensors is sent to a visualization module which graphically illustrates in a world map, by using glyphs, the main sentiment and the emotions arising from the people for a specific topic. This approach can help for deep understanding people's behavior and for providing at the same time a number of indicative factors regarding specific problems, people, ideas, items, etc.

2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) [11][15] is a paradigm to extract and characterize the meaning of words by statistical computations applied to a large corpus of texts. LSA is based on the vector space method: a text corpus is represented as a matrix $\mathbf{M}$ where rows are related to words, and columns are associated to documents or other contexts. The LSA paradigm defines a mapping between words and documents inducing from data a continuous vector space S, where the $i-th$ word w_i of a dictionary is associated to a vector $\mathbf{u}_i$ in S, and the $j-th$ document d_j of a text corpus is associated a vector $\mathbf{v}_j$ in S [20].

The S vector space is a "semantic space", since semantics conveyed by the presence of the i-th word in the j-th document can be measured by taking the dot product between the vector representing the word and the vector representing the document. The LSA approach learns automatically the similarity between the meanings of words, and bridges the gap between the information available in a set of text chunks and the knowledge people acquires after a large amount of experience. LSA has besides presented as a theory of learning, memory and knowledge; furthermore it is roughly equivalent to a neural network model, which is a typical bio-inspired cognitive architecture. As reported by Landauer et al, [12] it is supposed that the mind-brain stores and reprocesses its inputs in some manner that has the same effect of the Singular Value Decomposition operation.

3 The proposed approach

The proposed framework makes use of two kind of lexicons: *a)* a lexicon S containing a list of positive and negative subjective words from the Janyce Wiebe's subjectivity lexicon [17]; *b)* a lexicon E derived by Strapparava et al. in [18] containing words that are related to the six basic emotions of Ekman: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. The whole approach consists of two phases: a *training* phase and an *production* phase, which are illustrated below.

During the *training* phase the two lexicons are merged together in order to obtain a unique list of words. This list is then used by a module that, exploiting the Twitter APIs, retrieves from the Twitter stream those tweets that are written in english language and that contain a given word. The set of retrieved tweets constitutes a corpus, which is used to automatically induce the semantic space by means of LSA. Once the LSA space is built, it can be used to map both

the words that are present in the lexicons and the tweets that will be retrieved by the Twitter stream and that are object of the specific analysis illustrated below. For the experiment run in this paper we have used a set of 637841 unique tweets "sensed" and stored from the Twitter stream, choosing a truncation parameter $r = 300$ for the creation of the semantic space through LSA.

During the *production* phase the user makes a query with specific keywords regarding a topic of interest. Tweets are therefore retrieved and their textual content is mapped into the semantic space built during the training phase. Once the tweet text has been mapped as a vector $\mathbf{t}$, a k-nearest neighbour algorithm is run in order to catch the polarity and the emotions carried by tweet. In particular: let $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N$ the set of vectors representing the words in the union of the two lexicons S and E. To each $\mathbf{w}_i$ with $i = 1, 2, \cdots, N$ it is associated a 8-dimensional vector $\mathbf{v}_i = [v_{i1}, v_{i1}, \cdots, v_{i8}]$ whose components can be either 0 or 1. The first two dimensions are associated to the sentiment orientation; i.e., $v_{i1} = 1, v_{i2} = 0$ means "positive" orientation, while $v_{i1} = 0, v_{i2} = 1$ means "negative" orientation. The other 6 dimensions are related to the "emotions" pattern; e.g: $v_{i3} = 1, v_{i4} = 0, v_{i4} = 0, v_{i4} = 0, v_{i4} = 0, v_{i4} = 0$ is related to a "surprise" emotion. A k-NN algorithm is then run by computing the similarity between $\mathbf{t}$ and all the vectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N$. The similarity $sim_i(\mathbf{t}, \mathbf{w}_i)$ between $\mathbf{t}$ and $\mathbf{w}_i$, is computed as the cosine between them. The K vectors $\mathbf{w}_i$ most similar to $\mathbf{t}$ are then selected and the corresponding indexes i will constitute i set I. The following vector $\mathbf{p}$ is then computed: $\mathbf{p} = \sum_{i \in I} sim_i \mathbf{v}_i$, which represents the pattern of emotions and sentiments induced from the Latent Semantic Space. Starting from $\mathbf{p}$, two binary vectors $\overline{V_s}$ and $\overline{V_e}$ are computed.

The vector $\overline{V_s}$ is a two dimensional vector calculated by taking into account the components $[p_1, p_2]$ of $\mathbf{p}$: the $j-th$ component of $\overline{V_s}$ is set to 1 for the corresponding component in $[p_1, p_2]$ whose value is the highest; the other component is set to 0. If both components p_1 and p_2 have the same value, both the components in $\overline{V_s}$ are set to 0.

The computation of $\overline{V_e}$ is slightly different. Its dimensions are 7 and not 6 because we consider also a "neutral" emotion. For doing this, firstly, the average value p_{av} among the components $[p_3, p_4, p_5, p_6, p_7, p_8]$ is computed; then we set the i-th component of $\overline{V_e}$ to 1 for the highest value of p_i chosen among the p_i components of $[p_3, p_4, p_5, p_6, p_7, p_8]$ which are above the 20% of p_{av}; the other components are set to 0. If none of the components $[p_3, p_4, p_5, p_6, p_7, p_8]$ is higher then the 20% of p_{av} all the six components of $\overline{V_e}$ and its 0-th component is set to 1, indicating a "neutral" emotional pattern.

After the computation of $\overline{V_s}$ and $\overline{V_e}$, another 2-dimensional vector is built: it is intended to contain the GPS co-ordinates, if they are available, of the tweet taken into account; we named this vector as $\overline{V_c}$.

The three vectors $\overline{V_s}$, $\overline{V_s}$, and $\overline{V_s}$ can be seen as the results of three "soft sensors" each one sensing specific peculiarities of a tweet, namely the sentiment expressed in the text, the emotions arising from the tweet and the geospatial coordinates of the tweet. All these vectors, together with the text content of the tweet are then given to the visualisation module that exploits glyphs and colors to visually express the sensed properties from the Twitter stream.

4 Data Visualization

According to Friedman: *"Main goal of data visualization is to communicate information clearly and effectively through graphical means..."* [8]. An appropriate visualization of data makes the users able to understand and to reason about data also when they represent very complex phenomena. In our case we manage a very large number of values concerning the sentiments and the emotions related to a topic achieved by a sentiment analysis algorithm. Thus, an

effective visualization of this kind of data is a crucial aspect to bring down the complexity of the phenomena under investigation. We focused on the expressive power of colors, maps and glyphs combined all together.

The Data Visualization module of the framework acquires data coming from the soft sensors and exploits them in order to visualise data in a user friendly manner. In order to reach this goal, a specific color has been assigned to each kind of emotion. Considering the six base Ekman emotions plus a neutral emotion we have to manage seven emotions overall. Moreover, with regard to the sentiment polarity (positive or negative) we can get two integer values (zero or one). Thus, for each tweet we get a vector of eleven parameters coming from the soft sensors:

$$\overline{V} = [\overline{V_e}; \overline{V_s}; \overline{V_c}] = [\underbrace{emotions}_{\overline{V_e}7param.} | \underbrace{sentiments}_{\overline{V_s}2param.} | \underbrace{coordinates}_{\overline{V_c}2param.}] \quad (1)$$

The emotions sub vector $\overline{V_e}$ is a binary vector in which only one bit at time can be equal to one for each tweet. The position of the bit one, represents the dominant emotion associated to this tweet. Thus, a color is assigned analyzing the firsts seven parameters of the vector in $\overline{V}$ according to the following lookup table:

Look-up table of the colors assigned to the emotions							
$\overline{V_e}$	[1000000]	[0100000]	[0010000]	[0001000]	[0000100]	[0000010]	[0000001]
C	White	Cyan	Green	Yellow	Red	Blue	Violet
E	Neutral	Surprise	Fear	Joy	Anger	Sadness	Disgust

Table 1: Lookup table of the colors (C) assigned to the emotions (E) considering the vector $\overline{V_e}$

When geographical coordinates are available in a single tweet they are used to plot a placeholder on a planisphere map (see fig 1), which is an interactive map. Tweets which are geographically close are represented as a cluster with a superimposed number (see the orange circle of fig 1), this number indicates the quantity of tweets that the cluster represents. A zoom action on the map makes the clusters "explode" showing either single tweets or other sub clusters.

By using of this kind of data visualization an important target is achieved. In fact, it makes the users able to understand which countries are interested to a specific topic. In fact, just looking at the map, the geographical distribution of the tweets concerning a specific topic is immediately available. Moreover, the visual perception of the dominant color of the placeholder on the map, or in a specific region, makes clear the prevalent emotion associated to the topic under investigation. Finally, by combining the two previous concepts, the geographical distribution and the dominant color, users are able also to understand which emotions arise from a specific topic in different regions on the maps.

A similar map is built in order to represent the sentiment associated to a specific topic. In this case the color of the placeholder corresponding to a tweet can be black or white according to the values of the $\overline{V_s}$ sub vector. Also the $\overline{V_e}$ sub vector is a binary vector and also in this case only one bit at a time can be equal to one for each tweet. Thus, an high level (value 1) for the first bit represents a positive sentiment, otherwise an high level on the second bit represents a negative sentiment.

Taking into account a specific topic to investigate it is crucial to return an overall representation of the emotions and the sentiments associated with it. Thus, we need to represent in the same "graph" the composition of all sub vectors $\overline{V_e}$ and all sub vectors $\overline{V_s}$ associated to a specific topic. More in detail, by analyzing a topic, many tweets are associated to it: let this

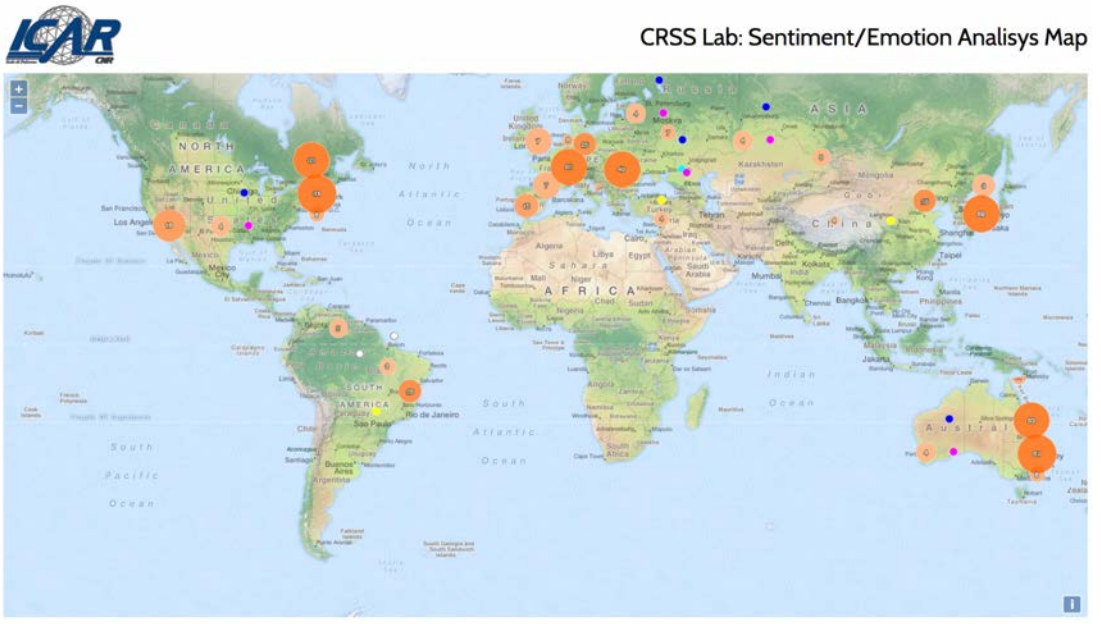

Figure 1: The tweets plotted as placeholders on a planisphere map. Orange circles with a superimposed number represent groups of tweets.

number be n, we get n; $\overline{V}$ and then n sub vectors $\overline{V_e}$ and n sub vectors $\overline{V_s}$. The composition of these n vectors can be computed as:

$$\overline{V_{Ce}} = \left(\frac{\sum_{k=1}^{n} \overline{V}_{ek}}{max(\sum_{k=1}^{n} \overline{V}_{ek})} \right) \quad \text{And} \quad \overline{V_{Cs}} = \left(\frac{\sum_{k=1}^{n} \overline{V}_{sk}}{max(\sum_{k=1}^{n} \overline{V}_{sk})} \right) \tag{2}$$

In doing so, two new vectors $\overline{V_{Ce}}$ and $\overline{V_{Cs}}$ are obtained, whose elements are real, subsequently they are recomputed so that just one element is one. Each element of $\overline{V_{Ce}}$ represents the percentage of the associated emotions and, in the same manner, each element of $\overline{V_{Cs}}$ represents the percentage of positiveness or negativeness.

The need is to find an effective representation of eight parameters (seven emotions and a real value representing a sentiment varying between total negativeness and total positiveness. Data visualization in this case must return an immediate perception of the distribution of the emotions related to a specific topic and at the same time the degree of positiveness or negativeness.

We have chosen a parametric curve called sinusoidal spiral properly modified and enriched with a circle. Colors, also in this case, play a fundamental role as will be shown in the following.

The sinusoidal spiral is a parametric curve studied by the Scottish mathematician Colin Maclaurin (1698-1746) for the first time. This curve, for some values of its parameters, resembles to a flower with several petals. The sinusoidal spiral is based on the polar equation formally defined as:

$$\rho^n = a^n * cos(n\theta) \quad \text{With } n \in \mathbb{R}. \tag{3}$$

Indeed, it is not a true spiral because its radius increases and decreases sinuously despite of it acts as a spiral. This curve can be thought as composed by rotation of a base pattern (see

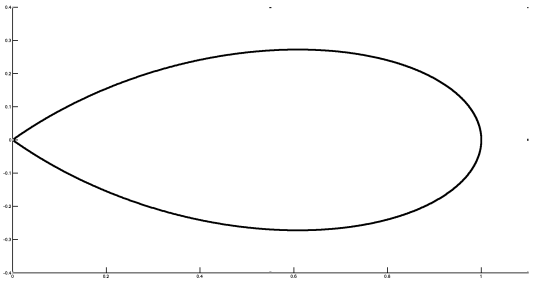

Figure 2: The base pattern of a sinusoidal spiral. The length of the petal is proportional to parameter a of equation (3). In this case $a = 1$

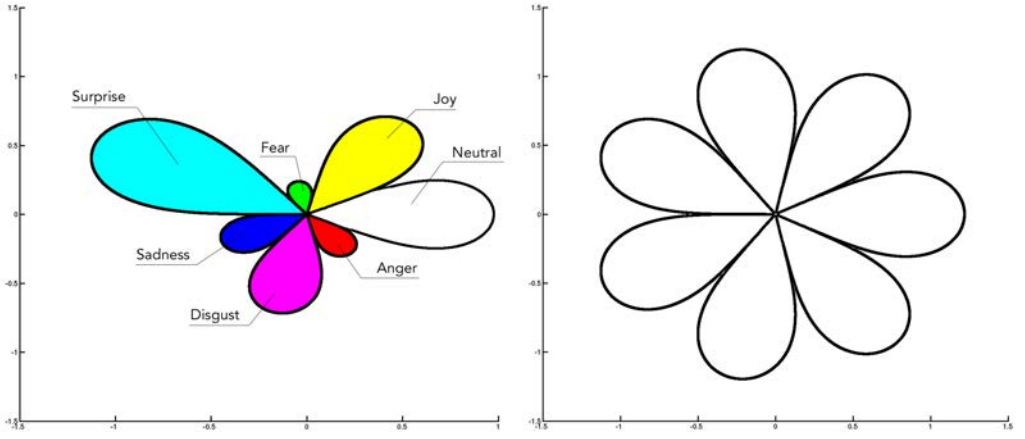

Figure 3: Right side: a void flower composed by seven petals. Left side: a filled flower composed by seven petals, the length of the petals is given by the values a_i in the vector $\overline{V_{Ce}} = [0.1, 0.4, 0.3, 0.7, 1, 0.6, 0.8]$. Each petal is filled with the corresponding color of the emotion.

figure 2) whose size can be controlled by a parameter of equation (3). The base pattern of figure 2 represents a petal of the flower and it is obtained according to equation (3) by varying θ in the range $[-\pi/2n, \pi/2n]$. The complete flower is obtained for all rotations of angle $2k\pi/n$ where $k \in \mathbb{N}$. If n is also a rational number, $n = p/q$ with $p, q \in \mathbb{N}$, the flower can be achieved performing $p-1$ rotations of the base pattern varying K in $[1, p-1]$. In our application we have chosen $q = 2$, so $n = p/2$. Now, choosing $p = 7$, a seven petals flower is defined. Each petal of this flower is related with an emotion and it is filled by a specific color according to Table 1. Moreover, the length (parameter a of equation 3) of a single petal (see the base pattern shown in fig 2) may be associated with a corresponding value in the vector $\overline{V_{Ce}}$. In doing so a single petal of this curve represents, at the same time, the kind of the emotions by its color and the percentage of its occurrence for a specific topic by its length, as shown on the left side of fig 3.

The flower obtained by the use of the sinusoidal spiral can be modified in order to represent, in the some glyph, the dominant sentiment related to a specific topic. To this aim, the basic pattern is shifted along the X axis. In doing so a circle, representing the pistil of the flower, can

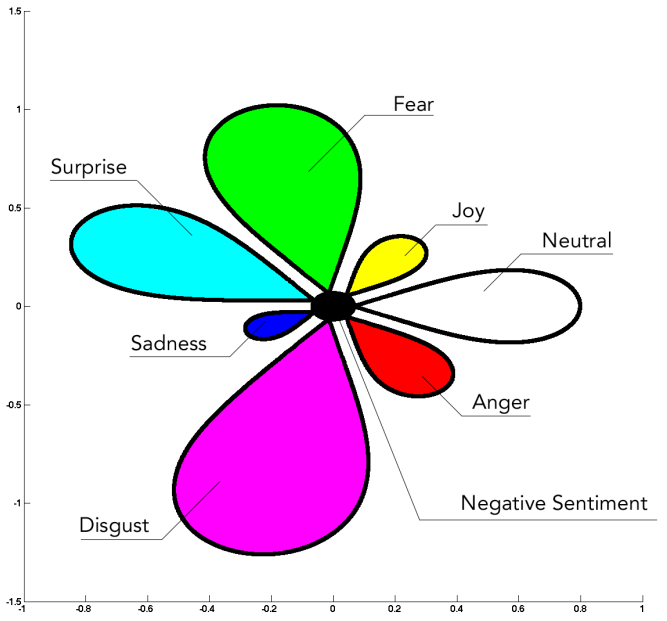

Figure 4: The flower glyph composed by seven petals, the length of the petals is given by the values a_i of the vector $\overline{V_{Ce}} = [0.8, 0.2, 0.4, 0.3, 0.7, 1, 0.6, 0.2]$. Each petal is filled with an emotoion color. The pistil represents the dominant sentiment. In this case it is black and represents a negative sentiment being the vector $\overline{V_{Cs}} = [1, 0]$.

be added among the petals. This circle can be filled by black solid color in order to represent a negative dominant sentiment, otherwise it can be filled by white solid color in order to represent a positive dominant sentiment. the dominant sentiment is achivied by the bits of the $\overline{V_{Cs}}$ sub vector (see fig 4.

5 Conclusions and Future Work

A preliminary work on the realization of a framework aimed at executing a social sensing task has been presented. The approach is founded on the Latent Semantic Analysis approach, which, thanks to its analogy with neural networks, can be seen as a bio-inspired cognitive architecture. Besides, LSA has the strength of having many interesting cognitive and statistical properties, being successfully used for simulating many human cognitive phenomena, and being used as a statistical estimator [15]. LSA is therefore the ideal mean for mapping tweets in a semantic space, where properly tuned soft sensors can detect and measure the emotional and sentimental content. This, joined with the spatial coordinates present in tweets, allows us to realize a glyph based interactive map, which can be very useful for exploring and analyzing in real time the emotional reactions of people on specific topics, goods, events, etc. In the future we are planning to test the approach with other techniques and to enhance both the retrieval and the soft sensors, in order to make the framework more affordable and effective.

References

[1] C. C. Aggarwal, T. Abdelzaher, "Integrating sensors and social networks" in Social Network Data Analytics, C. C. Aggarwal, Ed., Springer-Verlag, 2011, ch. 14, pp. 397–412.

[2] Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING), pages 36?44.

[3] Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. Computational Intelligence, 29(3).

[4] Canales, Lea and Martinez-Barco, Patricio (2014) Emotion Detection from text: A Survey. In: 11th International Workshop on Natural Language Processing and Cognitive Science - NAACL.

[5] D'Urso, V. and R. Trentin, 1998. Introduzione alla psicologia delle emozioni. Laterza.

[6] N. Eagle and A. Pentland. "Reality mining: Sensing complex social systems" Pers. Ubiquitous Comput., vol. 10, no. 4, pp. 255–268, 2006.

[7] Ekman, P. (1999). Basic Emotions. John Wiley and Sons Ltd, New York.

[8] Vitaly Friedman (2008) "Data Visualization and Infographics" in: Graphics, Monday Inspiration, January 14th, 2008

[9] A. Kanavos, I. Perikos, I. Hatzilygeroudis , A. Tsakalidis (2016) "Integrating User?s Emotional Behavior for Community Detection in Social Networks" Proc. of he 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 1, pages 355-362

[10] V.Krishnamurthy, H.V.Poor, (2014) "A Tutorial on Interactive Sensing in Social Networks" IEEE Trans on Computational Social Systems Vol1 No.1 March 2014, pp 3-21

[11] T.K. Landauer, P.W. Foltz, D. Laham. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, vol 25, pp.259-284.

[12] Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211–240.

[13] Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data, pages 415?463.

[14] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the International Conference on Language Resources and Evaluation (LREC).

[15] G. Pilato, G. Vassallo (2015) "TSVD as a Statistical Estimator in the Latent Semantic Analysis Paradigm". IEEE Trans. Emerging Topics Comput. 3(2): 185-192

[16] Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. (2012). Tracking ?gross community happiness? from tweets. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW ?12, pages 965?968.

[17] E. Riloff, J. Wiebe "Learning extraction patterns for subjective expressions" EMNLP 2003 *http://www.cs.pitt.edu/mpqa*

[18] C. Strapparava, A. Valitutti, "WordNet-Affect: an affective extension of WordNet??. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083–1086. *http://www.cse.unt.edu/ rada/affectivetext/*

[19] http://www.twitter.com

[20] D. Widdows, S. Cederberg, B. Dorow. 2002. Visualisation Techniques for Analysing Meaning. Fifth International Conference on Text, Speech and Dialogue. Brno, Czech Republic, September 2002. pp. 107-115.